

Characterizing humans on Riemannian manifolds

Diego Tosato, Mauro Spera, Marco Cristani, *Member, IEEE*, Vittorio Murino, *Senior Member, IEEE*.

Abstract—In surveillance applications, head and body orientation of people is of primary importance for assessing many behavioural traits. Unfortunately, in this context people is often encoded by few, noisy pixels, so that their characterization is difficult. We face this issue, proposing a computational framework which is based on an expressive descriptor, the covariance of features. Covariances have been employed for pedestrian detection purposes, actually, a binary classification problem on Riemannian manifolds. In this paper, we show how to extend to the multi-classification case, presenting a novel descriptor, named Weighted ARray of COvariances, WARCO, especially suited for dealing with tiny image representations. The extension requires a novel differential geometry approach, in which covariances are projected on a unique tangent space, where standard machine learning techniques can be applied. In particular, we adopt the Campbell-Baker-Hausdorff expansion as a means to approximate on the tangent space the genuine (geodesic) distances on the manifold, in a very efficient way. We test our methodology on multiple benchmark datasets, and also propose new testing sets, getting convincing results in all the cases.

Index Terms—Pedestrian characterization, Covariance descriptors, Riemannian manifolds.

1 INTRODUCTION

IN computer vision, and especially in videosurveillance, the capability of characterizing humans is surely of primary importance. In this regard, social signal processing studies [1] support the hypothesis that the body appearance is critical for inferring many behavioral traits, yielding to fine activity profiles. For example, head direction is fundamental for discovering the focus of attention of individuals [2], [3] and detecting interacting people [4], body posture and gestures during an interaction are typically indicators of speaking activity [5].

Characterizing humans becomes particularly troublesome whenever we handle small, noisy images. In such cases, tasks as body or head orientation estimation (see Fig. 1(a)) turn out to be serious challenges. This fact induced researchers to design novel features, such as robust classifiers or regressors, for exploiting the available small buch of pixels at best.

Recently, the use of covariance descriptors as composite features emerged as a powerful means for pedestrian detection [6]. In general, covariances showed to be naturally suited for encoding classes of objects with high intra-class variation, actually exploiting it for systematically encoding mutual relations among basic cues (as gradient, pixel intensity, etc.) [7]–[10]. For the

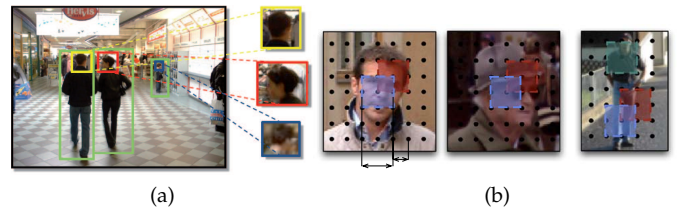


Fig. 1. (a) Example of an image from a video surveillance sequence containing pedestrians and close-up of their heads. (b) Weighted ARray of COvariance matrices (WARCO).

pedestrian case, Tuzel *et al.* [6] employed a boosting framework on Sym_d^+ , namely the set of positive definite $d \times d$ symmetric matrices (covariance matrices). The idea was to build weak learners by regression over the mappings of the training points on a suitable tangent space. This tangent space was defined over the weighted Karcher mean [11] of the positive training data points, so to preserve their local layout on Sym_d^+ . The negative points, instead, (i.e., all but pedestrians) were assumed to be spread on the manifold, without including them in the estimation of the mean.

In this paper, our aim is to move to a multi-class classification scenario, considering head and body orientations as object classes. In such a scenario, the above considerations do not hold any more, because we have many “positive” classes, each of them localized in a different part of the manifold. As a consequence, 1) choosing the Karcher mean of one class would privilege that class with respect to the others, and 2) the Karcher mean of all classes is inadequate. Therefore, our first contribution consists in a theoretical analysis of this

D. Tosato, M. Spera and M. Cristani are with the Dipartimento di Informatica, University of Verona, Strada le Grazie 15, 37134 Verona (Italy). Contacts: D. Tosato, e-mail diego.tosato@univr.it; M. Spera, e-mail mauro.spera@univr.it, Tel: +39 045 8027816; M. Cristani, e-mail marco.cristani@univr.it, Tel: +39 045 8027988, Fax: +39 045 8027068.

V. Murino is with the Istituto Italiano di Tecnologia, via Morego, 30, 16163 Genova and with the Dipartimento di Informatica, University of Verona, Strada le Grazie 15, 37134 Verona (Italy). Contacts: e-mail vittorio.murino@iit.it, Phone: +39 010 71781 504, Mobile: +39 329 6508554, Fax: +39 010 71781 236

space, so as to derive a point individuating a common *suitable* projection point that do not penalize any class. Such a point is chosen by analysing the local geometry of the manifold of the considered samples, realizing that, whenever the (sectional) curvature of the manifold is in general weak, a good candidate is the identity. This allows to consider covariance matrices as vectors in an Euclidean space where state-of-the-art classifiers can be utilized.

The second contribution consists in providing a novel measure for calculating distances between the projected points, in a way that the original geodesic distance is robustly preserved in a finer way with respect to the adoption of the Euclidean distance. This comes by considering the sectional curvature of the manifold, and adopting the general Campbell-Baker-Hausdorff (CBH) expansion [12].

In order to give a rough idea of it, and working with (square) matrices, CBH stems from the elementary fact that, since X and Y do not commute in general, one also has $\exp X \cdot \exp Y \neq \exp Y \cdot \exp X \neq \exp(X + Y)$. Hence, the CBH-formula, valid in any Lie algebra, is given as a series expansion in terms of nested commutators, of the following form:

$$\exp X \cdot \exp Y = \exp(X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] + \frac{1}{12}[Y, [Y, X]] + \dots), \quad (1)$$

where $[X, Y] = XY - YX$ (the standard matrix commutator). The CBH expansion allows us to detect the role of the curvature of the manifold, showing that the higher the curvature, the rougher the approximation of the distance. At the same time, our formulation provides a new approximation for the genuine geodesic distance on the manifold, finer than the Euclidean distance previously adopted [13]. We dubbed such an approximation CBH1, i.e., obtained by exploiting the first term of the CBH expansion.

As third contribution, we propose a novel object descriptor, expressively designed for encoding complex objects as pedestrians captured by few noisy pixels. The resulting descriptor is dubbed Weighted ARray of COvariances (WARCO) (see Fig. 1(b)), composed by a variable number of overlapped squared patches, each of them described by a covariance matrix of image features. Each covariance is fed into a local weighted classifier (a kernel classifier), where the weight - learned during the training stage - highlights its ability in encoding a defined portion of the object of interest. All the local classifiers are then linearly combined in a strong global classifier.

Adopting WARCO in the proposed theoretical framework allows to build robust kernel classifiers in a very economical way, since the building of the Gram-matrix turns out to be linear in the number of training examples as compared with the quadratic complexity in case of the (exact) geodesic distance.

A thorough experimental section on head orientation classification/regression and body orientation classification promotes our approach as a basic module for advanced surveillance, when fine analyses have to be carried out in difficult scenarios. In particular, we test on six different benchmark datasets (including QMUL head dataset, IDIAP head pose dataset, CAVIAR), proposing three novel sets for head and body orientation estimation. In all the cases, we get convincing results.

The rest of the paper is organized as follows. In Sec. 2, we report the related literature evidencing the novel aspects of our proposal. In Sec. 3, we present the mathematical analysis of Sym_d^+ , which has produced interesting theoretical findings exploited to design the statistical method. In Sec. 4, we describe the kernel-based classification model, which is extensively tested using several public datasets, whose results are illustrated in Sec. 5. Finally, conclusions and future perspectives are drawn in Sec. 6.

2 RELATED WORK

We focus our attention on models, object representations, and features for robust human body parts description and classification. In this context, the methods can be categorized in general-purpose, where the object models can be employed for different tasks, such as whole-body or human part detection, head and body orientation classification, (e.g. [13]–[17]) and task-specific models (e.g. [18]–[24]).

As for the task-specific models, we consider two tasks: head and body orientation classification. Prior to this, human detection methods should be also briefly reviewed, since the object model for the detection is usually inherited by head and body orientation classification procedures.

Typically, human detection approaches represent a human as a set of unsupervised selected parts [6], [18]–[20], [25]–[31], where such parts are represented by dense features such as Haar-wavelet-based descriptors, Shapelet [25], covariance matrices [6], part-templates [32], Joint Ranking of Granules (JRoG) [29], Local Binary Patterns (LBP) [26], [29], combination of HOG [18], Integral Channel Features [20], self-similarity on color channels [30], and synthesized features [31]. Other works combine some of the previously mentioned features as [29] where HOG and LBP are concatenated, and [27] where HOG, Haar-like, and Shapelet features are used.

Most of these approaches uses boosting for both a greedy estimation of the most discriminative patches and for classifying them at the same time. A relevant exception is [18], which presents a part-based deformable model for object detection. Considering HOG features [14], the object model is defined by a constellation of discriminative learned parts that score subwindows of a ROI (Region Of Interest) containing the OI, and the

classification framework is represented by latent Support Vector Machines (SVMs).

Considering now the head orientation estimation task, the literature is also vast [33]–[36]. For high resolution images, important methods are proposed in the context of the CLEAR07 challenge [34]. Instead, for low resolution images, the head orientation estimation task often translates into the head orientation classification task, in which there are few works in the state of the art. Two recent approaches [24], [37] provide valid solutions to these problems. Both works organize the overall processing scheme in two phases: detection and categorization.

Similarly to the head orientation estimation, for the human pose estimation task there are a consistent number of methods considering high-resolution images [23], [38]–[40]. Few methods deal with small pedestrians classifying their body orientation. An interesting example is [21], where a coarse-to-fine matching of an exemplar-based shape hierarchy and Chamfer distance are used to find the best template describing a candidate human orientation.

Considering general purpose models, probably the most important example is the detector proposed by Dalal & Triggs [14]. This detector, which uses as feature the HOG, still represents an effective solution to the object detection and classification tasks. HOG describes an object as a fine set of overlapping blocks and the algorithms utilize a sliding window procedure, where a discriminative SVM model is applied to all positions and scales of an image. This approach has been also used in [38] for human pose estimation, where the pose is recovered by direct regression of the HOG descriptors. Agarwal & Triggs have also demonstrated an application of non-negative matrix factorization that allows us to discriminate features of interest from background. Another interesting approach based on HOG features is proposed by Lin & Davis [32]. It adopts an OI model similar to the one we propose. In fact, instead of standard concatenation-style image location-based feature encoding, patches are evaluated independently and then a probabilistic framework is used to link the evaluation results. Some years later another successful work is proposed by Schwartz et al. [41]. It again uses HOG features on both color and gray scale images, and the pre-process the feature space using partial least squares to reduce its dimensionality.

Recently, in [42], the HOG representation was employed to categorize the pedestrian orientations in a few classes, considering pedestrians at low resolution. Moreover, HOG is in this case combined to adaptive local receptive field features in a multi-layer neural network architecture.

In [15], a different kind of histogram-based representation is used, based on the spatial pyramid concept [43]. These two models generalize the previous ones because a multi-layer analysis is performed, but a regular grid structure is still used to represent the object.

A different approach is used in [16], where patches are sampled randomly from images to build the object class model using Hough Forest, which is a Random Forest that directly maps the image patch appearance to the probabilistic vote about the possible location of the object centroid, similarly to the implicit shape model. Since fixed-size patches are used, the method is adaptable to a wide range of tasks.

The type of OI basic descriptors presented in the current work, i.e., covariance matrices, has been already exploited in the case of pedestrian detection [6], [9], tracking and retrieval [44] and in the biomedical research domain [45], [46]. A mathematical treatment of covariance matrices in Computer Vision is reported in [47], but the investigation of the properties of covariance matrices as objects living in a non Euclidean space is still an active research topic thanks to their versatility and effectiveness when used as descriptors for classification tasks [48]–[50].

The approach proposed here can be categorized as general purpose, and a former version was presented in [13]. The approach proposed here differs in several ways as a new weighted covariance descriptor is introduced, which is then exploited adopting a kernel machine architecture suitable for both classification and regression tasks. Moreover, the theoretical part is consistently new as we present a rigorous and comprehensive mathematical analysis of the covariance matrices living in a Riemannian manifold, whose findings are utilized to justify and lay down the ground of the proposed statistical classification method.

3 THEORETICAL ANALYSIS OF Sym_d^+

In this section, we aim at gathering together basic differential geometry notions about Sym_d^+ , namely the set of positive definite $d \times d$ symmetric matrices (covariance matrices), adopting the formalism of [12], [51]; our coverage will allow to introduce the main theoretical contribution of this work, that is, the application of the Campbell-Baker-Hausdorff expansion as a fast way to approximate distances in Sym_d^+ .

In particular, after recalling some preliminary notions in Sec. 3.1, we show that Sym_d^+ is a homogeneous space (Lemma 1): this means that we are entitled to select any point on Sym_d^+ for defining a tangent space over which projecting points and in which distances are calculated. In Sec. 3.2, we show that the identity I_d on Sym_d^+ is a particularly convenient choice (under a pure computational complexity aspect) as a projection point. In Sec. 3.3, we introduce the (sectional) curvature of Sym_d^+ which allows to measure how much Sym_d^+ differs from a Euclidean space, which is flat. In particular, we show here that Sym_d^+ has negative curvature (Lemma 2), and this will for instance ensure that there is only one geodesic connecting any two points; moreover, this will

show that the first correction to the Euclidean distance provided by the CBH-expansion, that is, our distance approximation, is non negative. This is finally discussed in Sec. 3.4, with Theorem 1.

3.1 Preliminaries

In general, given a Lie group G and a closed Lie subgroup H thereof, the quotient set G/H consisting of all left cosets $[g] := gH = \{gh \mid h \in H\}$ becomes in a unique way a smooth manifold (this is the prototype of a G -homogeneous space). The study of the geometrical properties of homogeneous spaces is greatly eased by the fact that all points can be treated on the same footing (colloquially, the manifold appears to be the same when looked upon from whatever point therein). This is quite important from the machine learning point of view. Therefore, it is natural, for both theoretical and practical reasons, to focus the attention on the class $[e] = H$ of the neutral element $e \in G$. A graphical example of homogeneous space is shown in Fig. 2(a).

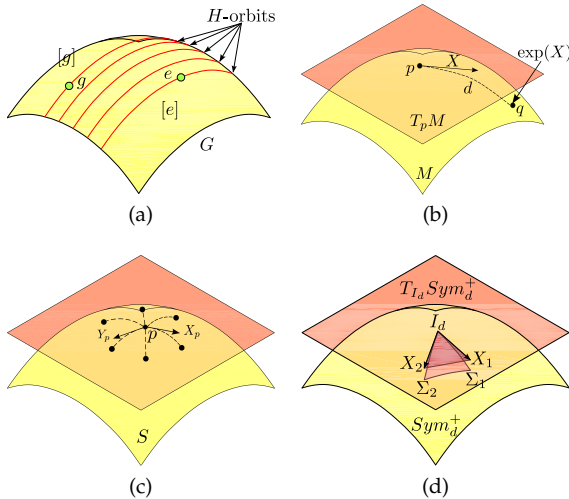


Fig. 2. (a) Homogeneous spaces. (b) Exponential map. (c) Gaussian curvature ($\kappa_p(X_p, Y_p)$) of the 2-dimensional surface S at p . (d) Approximating the true geodesic distance.

Lemma 1: Sym_d^+ is an homogeneous space.

Proof: The general linear group $Gl(d, \mathbb{R})$, consisting of all non-singular real $d \times d$ matrices, naturally acts on Sym_d^+ via congruence: $Sym_d^+ \ni \Sigma \mapsto M^T \Sigma M \in Sym_d^+, M \in Gl(d, \mathbb{R})$. By virtue of (a corollary of) Sylvester's theorem [52], the latter action is transitive: in other words, any two positive definite symmetric matrices are congruent, i.e., there is always an M that connects them. In particular, every matrix $\Sigma \in Sym_d^+$ is congruent to I_d (the $d \times d$ identity matrix): $\Sigma = M^T \cdot I_d \cdot M = M^T M$ for some $M \in Gl(d, \mathbb{R})$; in our scenario we shall take, for specific calculations, $M = \Sigma^{\frac{1}{2}}$. Therefore, Sym_d^+ is the space of all symmetric matrices congruent to I_d .

Also, I_d is invariant under congruence, namely $M^T M = I_d$, if and only if $M \in O(d, \mathbb{R})$, the group of orthogonal $d \times d$ matrices. In other words, $O(d, \mathbb{R})$ is the isotropy group of I_d . From this, one finds that Sym_d^+ is the homogeneous space $Sym_d^+ \cong Gl(d, \mathbb{R})/O(d, \mathbb{R}) \cong Gl_+(d, \mathbb{R})/SO(d, \mathbb{R})$ (one may restrict to matrices with positive determinant to get connected groups). $SO(d, \mathbb{R})$ denotes the special orthogonal group, i.e. the orthogonal matrices having determinant +1. \square

In view of the homogeneity, we choose to work at the identity, since this will ease all subsequent computations.

3.2 A Riemannian metric on Sym_d^+

Recall that a Riemannian manifold $(\mathcal{M}, \langle \cdot, \cdot \rangle)$ is a smooth manifold equipped with a Riemannian metric $\langle \cdot, \cdot \rangle$, i.e. a smoothly varying inner product $\langle \cdot, \cdot \rangle_p$ on its tangent spaces $T_p \mathcal{M}$, $p \in \mathcal{M}$. The tangent vectors (the elements of $T_p \mathcal{M}$) are the “velocities” of the curves in \mathcal{M} issuing from $p \in \mathcal{M}$ or, equivalently, the “directional derivatives” of the smooth functions defined in a neighbourhood of p .

The tangent space of Sym_d^+ at any point Σ (notation: $T_\Sigma Sym_d^+$), is Sym_d , the space of symmetric matrices. By homogeneity it is enough to check this at the identity I_d . Indeed, let us consider an interval $J \subset \mathbb{R}$ containing 0, and let us consider a smooth curve of matrices $J \ni t \mapsto \Sigma(t) \in Sym_d^+$ with $\Sigma(0) = I_d$. Its “velocity” at I_d , namely $\dot{\Sigma}(0)$, belongs to Sym_d , since the derivative of $\Sigma(t)$ is still a symmetric matrix. Vice versa, given a matrix $W \in Sym_d$, it is possible to find a curve in Sym_d^+ starting at I_d with velocity given by $W = \dot{\Sigma}(0)$. Taking for instance $\Sigma(t) = \exp(tW)$, if we diagonalize the matrix W and denote its eigenvalues by w_i , $i = 1, 2, \dots, d$, then the eigenvalues of $\Sigma(t)$ are $\exp(tw_i) > 0$, $i = 1, 2, \dots, d$. Therefore, the matrix is positive definite. By continuity, any curve with the same velocity at I_d is locally in Sym_d^+ . Given $\varphi \equiv \varphi_M : Sym_d^+ \ni \Sigma \mapsto M^T \Sigma M \in Sym_d^+$, its differential φ_* is:

$$\varphi_* : T_{I_d} Sym_d^+ \rightarrow T_\Sigma Sym_d^+, \quad W' \mapsto M^T W' M =: W, \quad (2)$$

since in general $D(M^T \Sigma M) = M^T D \Sigma M$ (here D denotes of course differentiation; notice that W is symmetric as well, as asserted before).

On $T_{I_d} Sym_d^+$ it is possible to define the Frobenius inner product:

$$\langle W'_1, W'_2 \rangle_{I_d} := Tr(W'_1 W'_2), \quad (3)$$

where $W'_i \in Sym_d$. It is extended to a Riemannian metric, invariant under congruence via the formula:

$$\langle W_1, W_2 \rangle_{\varphi(I_d) = \Sigma} := \langle \varphi_*^{-1}(W_1), \varphi_*^{-1}(W_2) \rangle_{I_d}, \quad (4)$$

namely (by a short computation using $\Sigma = M^T M$)

$$\langle W_1, W_2 \rangle_\Sigma = Tr(\Sigma^{-1} W_1 \Sigma^{-1} W_2), \quad (5)$$

where $W_1, W_2 \in T_\Sigma Sym_d^+ \cong Sym_d$. This turns out to be well-defined since the Frobenius inner product is

$O(d, \mathbb{R})$ -invariant: indeed, if $O \in O(d, \mathbb{R})$, we have:

$$\begin{aligned} \text{Tr}(O^T W_1' O \cdot O^T W_2' O) &= \text{Tr}(O^T W_1' W_2' O) \\ &= \text{Tr}(W_1' W_2' O O^T) \\ &= \text{Tr}(W_1' W_2') \end{aligned} \quad (6)$$

This further entails that, given any two points $\Sigma_1, \Sigma_2 \in \text{Sym}_d^+$, and $\varphi \equiv \varphi_M : \text{Sym}_d^+ \ni \Sigma \mapsto M^T \Sigma M \in \text{Sym}_d^+$, then

$$d(\varphi(\Sigma_1), \varphi(\Sigma_2)) = d(\Sigma_1, \Sigma_2), \quad (7)$$

where d is the distance induced by the above Riemannian metric (and equals the length of a minimal geodesic connecting the two points - in our case the latter exists and it is unique, see also below); in other words, φ is an *isometry*.

In particular, we may compute all distances from a fixed point, the natural choice thereof being the identity. Also, any $\Sigma \in \text{Sym}_d^+$ is of the form

$$\Sigma = \exp W_\Sigma, \quad W_\Sigma = \log \Sigma \in \text{Sym}_d \quad (8)$$

(spectral theorem), therefore $d^2(I_d, \Sigma) = \|\log \Sigma\|^2 = \text{Tr}((\log \Sigma)^2) = \sum_{i=1}^d (\log \sigma_i)^2$.

The σ_i 's are the (positive) eigenvalues of Σ and, in general (setting below $M_1^T M_1 = \Sigma_1$, and specifically $M_1 = \Sigma_1^{\frac{1}{2}}$):

$$\begin{aligned} d^2(\Sigma_1, \Sigma_2) &= d^2(\varphi_{M_1}(I_d), \Sigma_2) \\ &= d^2(I_d, \varphi_{M_1}^{-1}(\Sigma_2)) \\ &= \text{Tr}((\log(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}))^2) \\ &= \sum_{i=1}^d (\log \xi_i)^2 \end{aligned} \quad (9)$$

where the ξ_i 's are the (positive) eigenvalues of $\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}$. In fact Sym_d^+ is actually a *Riemannian symmetric space* $(\mathcal{M}, \langle \cdot, \cdot \rangle)$, namely, for each point $p \in \mathcal{M}$, there exists an isometry σ_p fulfilling $\sigma_p^2 = \text{Id}_{\mathcal{M}}$ (with $\text{Id}_{\mathcal{M}}$ the trivial isometry on \mathcal{M}) and having p as an isolated fixed point ([51]). We shall not delve any further into the general theory of symmetric spaces, confining ourselves to recalling specific facts when needed. For example, it follows from it that the geodesics starting from I_d are of the form

$$\mathbb{R} \ni t \mapsto \exp(tW) \in \text{Sym}_d^+, \quad W \in \text{Sym}_d, \quad (10)$$

with \exp the standard matrix exponential (since, for symmetric spaces associated to matrix groups, the Riemannian exponential coincides, at the identity, with the matrix one). An intuitive pictorial idea of the exponential map is illustrated in Fig. 2(b) In our case, the isometry σ_p of the general theory is induced at I_d by the map $\text{Sym}_d \ni W \mapsto -W \in \text{Sym}_d$.

3.3 Non-positivity of the sectional curvature of Sym_d^+

Given a Riemannian manifold $(\mathcal{M}, \langle \cdot, \cdot \rangle)$ its *sectional curvature* $\kappa_p(X_p, Y_p)$ at $p \in \mathcal{M}$, if X_p and Y_p are linearly independent tangent vectors at p , is given by

$$\kappa_p(X_p, Y_p) := \frac{\langle R(X_p, Y_p)X_p, Y_p \rangle_p}{\langle X_p, X_p \rangle_p \langle Y_p, Y_p \rangle_p - \langle X_p, Y_p \rangle_p^2} \quad (11)$$

where R is denoting the Riemann curvature operator (see below). Notice that the denominator represents the area squared of the parallelogram determined by X_p and Y_p .

It is important to pinpoint that the sectional curvature just depends on the plane spanned by X_p and Y_p , and indeed it turns out to coincide with the Gaussian curvature, at p , of the parametric surface $S : (u, v) \mapsto \exp_p(uX_p + vY_p)$ (here \exp_p denotes the Riemannian exponential at p). We show an example of that in Fig. 2(c). In geometry, the (sectional) curvature is a measure of non-flatness of the manifold. The local vanishing of the curvature implies that the Riemannian manifold in question is actually a portion of a Euclidean space. We shall exploit this for learning purposes.

Lemma 2: The sectional curvature for Sym_d^+ is non-positive at any point.

Proof: Since Sym_d^+ is a symmetric space, one can again work at the identity, whereat one gets the following expression for the Riemann curvature operator (in the symmetric space framework, see e.g. [51])

$$R(X, Y) : \text{Sym}_d \ni Z \mapsto [[X, Y], Z] \in \text{Sym}_d. \quad (12)$$

Here, $[X, Y] = XY - YX$ is the matrix commutator. Then the sectional curvature κ_{I_d} at I_d reads (with $X, Y \in \text{Sym}_d$ linearly independent):

$$\begin{aligned} \kappa_{I_d}(X, Y) &= \frac{\langle R(X, Y)X, Y \rangle}{\|X\|^2 \|Y\|^2 - \langle X, Y \rangle^2} \\ &= 2 \frac{\text{Tr}((XY)^2 - X^2 Y^2)}{\text{Tr}(X^2) \text{Tr}(Y^2) - (\text{Tr}(XY))^2}, \end{aligned} \quad (13)$$

by the cyclical property of the trace.

Again, the denominator $\|X_1\|^2 \|X_2\|^2 - \langle X_1, X_2 \rangle^2 =: \mathcal{A}(X_1, X_2)^2$ is the area of the parallelogram determined by X_1 and X_2 , squared. Therefore, to prove that $\kappa_{I_d}(X, Y) \leq 0$, it suffices to show that

$$\text{Tr}((XY)^2) \leq \text{Tr}(X^2 Y^2), \quad (14)$$

and that equality holds if and only if $[X, Y] = 0$. This is implied by the following immediate consequence of the Schwarz inequality for (real) inner products

$$\langle x, y \rangle \leq \|x\| \|y\|, \quad \text{if } \|x\| = \|y\| \quad (15)$$

(equality holding if and only if $x = y$). Indeed, upon setting $x = XY$, $y = YX$, $\langle x, y \rangle = \text{Tr}(x^T y) = \text{Tr}(y^T x)$,

and using $X^T = X$, $Y^T = Y$, we have:

$$\begin{aligned}\|y\|^2 &= \text{Tr}((YX)^T(YX)) \\ &= \text{Tr}(X^T Y^T Y X) \\ &= \text{Tr}(XY Y X) \\ &= \text{Tr}(X^2 Y^2) = \|x\|^2\end{aligned}\quad (16)$$

□

We again stress that, for learning purposes, $\kappa_p(X_p, Y_p)$ provides a quantitative measure of how much a Riemannian manifold differs from a flat (i.e. Euclidean) one.

3.4 An expansion of the distance via the CBH-formula

Recalling that Preismann's theorem (see e.g. [51]) says that any two points of a complete simply connected manifold with non-positive sectional curvature are connected by precisely one geodesic, one has that, given a geodesic triangle with sides of length a, b, c , and angle θ opposite to (the side with length) c , the $a^2 + b^2 - 2ab \cos \theta \leq c^2$ inequality holds. An application of the theorem to Sym_d^+ (which indeed satisfies the above assumptions) shows that, taking the geodesic triangle with vertices I_d, Σ_1, Σ_2 , one gets $d_{\mathcal{E}}(\log_{I_d} \Sigma_1, \log_{I_d} \Sigma_2) \leq d(\Sigma_1, \Sigma_2)$, where $d_{\mathcal{E}}$ denotes the standard Euclidean distance (induced by the Frobenius norm)

$$d_{\mathcal{E}}^2(X_1, X_2) = \text{Tr}((X_1 - X_2)^2) \quad (17)$$

with $X_i = \log_{I_d} \Sigma_i$. But actually one can easily get approximate formulae for the distance by exploiting the Campbell-Baker-Hausdorff formula (CBH) (see e.g. [12], p.30, where the more general Dynkin's formula is given, and below: we shall apply it to the Lie algebra consisting of real $d \times d$ matrices).

Namely, we are going to show the following:

Theorem 1: The crudest approximation beyond the Euclidean distance (computed on the tangent space $T_{I_d} Sym_d^+$; also set $X_i := \log \Sigma_i$, $i = 1, 2$) reads:

$$\begin{aligned}d^2(\Sigma_1, \Sigma_2) &= d_{\mathcal{E}}^2(X_1, X_2) - \frac{1}{12} \langle R(X_1, X_2) X_1, X_2 \rangle + \dots \\ &= d_{\mathcal{E}}^2(X_1, X_2) - \frac{1}{12} \kappa(X_1, X_2) \cdot \mathcal{A}(X_1, X_2)^2 + \dots\end{aligned}\quad (18)$$

(which we illustrate in Fig. 2(d).)

Proof: The calculation employs the CBH-formula (suitably truncated to second order commutators)

$$\begin{aligned}\log(e^X e^Y) &= \\ &= X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] + \frac{1}{12}[Y, [Y, X]] + \dots\end{aligned}\quad (19)$$

which subsequently entails

$$\log(e^X e^Y e^X) = 2X + Y - \frac{1}{6}[X, [X, Y]] - \frac{1}{6}[Y, [X, Y]] + \dots \quad (20)$$

The above series are indeed convergent. Upon setting $X = -\frac{1}{2}X_1$, $Y = X_2$, the r.h.s. of the above formula becomes

$$W = X_2 - X_1 - \frac{1}{24}[X_1, [X_1, X_2]] + \frac{1}{12}[X_2, [X_1, X_2]] + \dots \quad (21)$$

Now, substituting the above expression in the formula for the distance (Eq. (9)), we find, after a short computation exploiting the properties of Tr :

$$\begin{aligned}d^2(\Sigma_1, \Sigma_2) &= \\ &= \text{Tr}[(X_2 - X_1)^2] - \frac{1}{12} \text{Tr}\{[X_1, [X_1, X_2]](X_2 - X_1)\} \\ &\quad + \frac{1}{6} \text{Tr}\{[X_2, [X_1, X_2]](X_2 - X_1)\} + \dots\end{aligned}\quad (22)$$

The last expression can be eventually transformed into Eq. (18) upon recalling the formula for the Riemannian curvature operator (Eq. (12)), together with the following general Riemann tensor identities (the third one being the *Bianchi identity*, see e.g. [51]):

$$R(x, y, z, t) = -R(y, x, z, t) = -R(x, y, t, z) = R(z, t, x, y) \quad (23)$$

$$R(x, y, z, t) + R(y, z, x, t) + R(z, x, y, t) = 0 \quad (24)$$

where $R(x, y, z, t)$ is defined as $\langle R(x, y)z, t \rangle$. In particular, we have

$$R(X_1, X_2, X_1, X_1) = R(X_1, X_2, X_2, X_2) = 0, \quad (25)$$

and we easily get the sought-for approximate formula (18). □

In Sec. 5 we will show the efficacy of the expansion above in approximating the geodesic distance ¹.

4 THE STATISTICAL FRAMEWORK

4.1 The General Architecture

The WARCO classifier has been designed specifically to deal with few visual information, that is, tiny images with noisy pixel values. It consists in a grid of N_p uniformly spaced and overlapped $k \times k$ patches $\Phi = \{\phi_n\}_{n=1, \dots, N_p}$, where each patch is described by a covariance matrix of features. For the sake of generality, we do not specify here neither the degree of overlap nor the nature of the feature considered, postponing this aim in the experiments.

In a L -class classification scenario, ARCO instantiates a single classifier on each patch, and provides a posterior classification probability which is

$$P(l|\Phi) = P(l|\{\phi_n\}_{n=1, \dots, N_p}) = \sum_{n=1}^{N_p} w_n P(l|\phi_n). \quad (26)$$

1. We just kept the first correction to the Euclidean distance. One could work out more refined expressions upon carefully keeping track of the various summands of CBH expansion. The successive terms, depending on nested commutators, are also related to curvature. Notice that we did not provide precise estimates for the approximation error.

where $l = 1, \dots, L$ is the class index, $n = 1, \dots, N_P$ is the patch index, $P(l|\phi_n)$ is the per-patch posterior probability of the n -th classifier and w_n is a normalized weight so that $\sum_n w_n = 1$, one for each patch classifier. This formulation is inspired by the Mixed Memory models (MMM) [53], that applies when a random variable is conditioned on the joint occurrence of a set of events; since the modeling of the joint conditional could be hard (due to complex dependencies between the variables, for example), the MMM approximates the joint conditional as a convex combination of pairwise conditionals. In our case, Eq. (26) approximates the fact that the classifiers are not independent (they actually work on local patches that in general are overlapped). The weights are learned in the following way: a 10-fold cross-validation strategy extracts a validation set; all the classifiers are trained on the remaining training set. On the validation set, all the classifiers give their votes; counting and normalizing the times the classifiers have done the correct choice gives a temporary weight. Averaging the temporary weights of all the runs of the cross-validation gives the final weights.

In a regression scenario, WARCO instantiates a regressor for each patch, and the final output is the median of all the outputs of the single regressors.

Standard Support Vector Machine (SVM) is the tool employed for performing classification and regression, where the Gram-matrix has been calculated by employing three different distances, i.e.,

$d_{\mathcal{E}}$: The distance between covariance matrices based on the Frobenius norm (see Sec. 3.2)

$$d_{\mathcal{E}}^2(X, Y) = \text{Tr}((\log_{I_d}(X) - \log_{I_d}(Y))^2). \quad (27)$$

d_{CBH1} : The distance² between covariance matrices exploiting the CBH expansion limited to the first order (see Sec. 3.4)

$$d_{\text{CBH1}}^2(X, Y) = d_{\mathcal{E}}^2(X, Y) + \tilde{\Xi}(\kappa_{I_d}), \quad (28)$$

where

$$\begin{aligned} \tilde{\Xi}(\kappa_{I_d}) = & \\ & - \frac{1}{12} \langle R(\log_{I_d}(X), \log_{I_d}(Y)) \log_{I_d}(X), \log_{I_d}(Y) \rangle. \end{aligned} \quad (29)$$

$d_{\mathcal{G}}$: The actual geodesic distance between covariance matrices (see Sec. 3.2)

$$d_{\mathcal{G}}^2(X, Y) = \text{Tr}(\log_{I_d}^2(X^{-\frac{1}{2}} Y X^{-\frac{1}{2}})). \quad (30)$$

We compute these three distances between all the training samples; therefore, for each distance, we obtain a dissimilarity matrix D . To use this distance measure in

2. Actually, we did not check whether d_{CBH1} is actually a distance in a rigorous mathematical sense. It is indeed symmetric, positive, and zero if and only if the points coincide, but one should further prove that it fulfils the triangle inequality; however, for our comparison purposes, we can safely call it, informally, distance.

Support Vector Machines, we use the extended Gaussian Kernel [54]: this amounts to apply the nonlinear transformation $\tilde{D} = \exp\left(\frac{-D}{\mu(D)}\right)$, where $\mu(D)$ is the average value of all the elements of D , making the resulting matrix a Mercer kernel.

Looking at the three distances, one can easily imagine the complexity for building the related Gram matrices. The logarithmic projection is without doubts the most demanding operation, so that it represents the bottleneck of the framework. Using the distances (27) and (28), the number of logarithmic projections is linear: in particular, in the CBH1 distance, all the N elements into play have to be projected over the identity, and then the projections can be employed for building the Gram matrix. Conversely, with the geodesic distance, all the elements have to be projected on the tangent spaces of all the elements, so that the complexity results quadratic. To give an practical intuition, whereas the learning of a classifier employing CBH1 with 10K elements takes 24 hours, considering the geodesic distance this translates in 576 hours (one month) on a Quad Core Intel Xeon Processor E5603 platform (1.6 GHz).

In the literature, there is a recent study [50] which shares some aspects with our approach; it proposes a fast and effective (dis)similarity function for Sym_d^+ using Jensen-Bregman (J-B) LogDet divergence, which also provides an associated fast search tree structure. This method is also shown to be much faster than the Riemannian metric, while being more accurate than the log-euclidean metric. In our case, our aim is to build kernels for SVM with N elements of dimension d . With CBH1 the complexity is governed by the matrix multiplication which is $O(d^2)$, whereas it is $O(d^3)$ for the Jensen-Bregman divergence (due to the determinants' computation). Moreover, although a closer scrutiny of the mutual relationship between the J-B LogDet approach and ours would be desirable, the former method appears to be rather ad hoc. Our proposed approach, instead, which is based on the intrinsic geometry of the manifold, can in principle be extended to other situations where Riemannian geometry can be applied.

4.2 Features

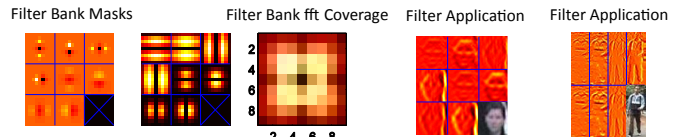


Fig. 3. On the left Symmetric DOOG (Difference Of Offset Gaussian) filters used to populate the feature set Φ . On the right two examples of their application on an head and a human image.

In our approach, we extract from each image I ($r \times r$ pixels), a set $\Phi(I, x, y)$ of dimension $r \times r \times d$ features

where $d = 13$ and x, y are the pixel location. It is composed by:

$$\Phi(I, x, y) = [F_1(Y) \dots F_8(Y) \ Y \ C_b \ C_r \ G_{| \cdot |}(Y) \ G_O(Y)], \quad (31)$$

where $F_1(Y) \dots F_8(Y)$ is the filter bank, depicted in Fig. 3, consisting of scaled symmetric DOOG (Difference Of Offset Gaussian) [55], applied only on the luminance channel of the perceptually uniform CIE Lab color space. Y , C_b , and C_r are the three color channels obtained by transforming the original *RGB* image. $G_{| \cdot |}(Y)$ and $G_O(Y)$ are the gradient magnitude and orientation calculated on the Y channel map, respectively.

5 EXPERIMENTS

In this section, we extensively test our approach in different tasks related to video surveillance.

In particular, we perform head orientation analysis in Sec. 5.3 and 5.4, and human orientation classification in Sec. 5.5. We do this under different operative conditions, with in total 6 datasets, each of them bringing in different issues. We also compare our proposal with known competitors, showing convincing performances.

In Table 1, first 5 columns, we summarize the nature of the datasets considered and the settings of WARCO (number of patches, size of the patches). In all the experiments, the overlap among patches is 0.5, since this value it has been empirically demonstrated a convenient choice [14].

On these datasets, we show how facts and intuitions of Sec. 3 can be clearly observed (Sec. 5.2). More specifically, we analyze the sectional curvature κ_{Id} , showing empirically that there is a relation between size of the patch, sectional curvature and approximation error of the Frobenius distance d_E (27) and CBH1 distance d_{CBH1} (28) (Sec. 3.3). We also demonstrate that, in average, $d_E \leq d_{CBH1} \leq d_G$ (see Sec. 3.4).

5.1 Datasets

For head orientation classification, we consider the *QMUL*, the *Heads Of CoffeeBreak* (HOCoffee), and the *Heads of IIT* (HIIT) [56] datasets. All the datasets are partitioned into a train and test set.

The *QMUL* head dataset (see Fig. 4(d) for some examples) is formed by head images taken from the *i-LIDS* dataset³ portraying an airport indoor scenario. It is composed by 19292 images: 10517 for the training and 8775 for the testing phase. They are uniformly partitioned into 5 classes: Back (BA), Front (FR), Left (LE), Right (RI), and Background (BG). Background images contain portions of the background scene.

The images are 50×50 pixels. The best performances are achieved in [13] in this case. The challenges of this dataset consist in scarce/non-homogeneous illumination, and quite severe occlusions.

The HOCoffee dataset (see Fig. 4(b)) is a novel benchmark dataset extracted from the CoffeeBreak social signal processing dataset [4], where an outdoor coffee break session during a summer school was captured, for detecting automatically social interactions. It is composed by 18117 head examples, 9522 in the training and 8595 in the testing set, of 50×50 pixels, uniformly partitioned into 6 different classes (orientations): Back, Front, Front-Left, Front-Right, Left, and Right. The images contain a margin of 10 pixels on average, so the actual average dimension of the heads is 30×30 pixels. HOCoffee images show two main issues: the heads are captured automatically by a head detector, therefore they are often not centered in the images. In addition, there are several important occlusions.

The HIIT dataset (see Fig. 4(a)) has been built combining some indoor image data captured in a controlled scenario (a vision lab) and the Pointing04 [57], Multi-PIE [58], and QMUL [56] datasets. As the previous dataset, it has 6 classes, 2000 examples each both for the training and testing set. The size of the samples is 50×50 pixels, without margin around the heads. The main characteristic of this dataset is that it has a stable background and no occlusions, so that it represents the ideal scenario where to evaluate how well a classifier can perform at a given resolution.

The *QMUL* and the HIIT dataset contain the images of the head of thousand of different subjects, while the HOCoffee focuses on 15 subjects taken in two different experimental sessions.

Considering the head orientation estimation, we focus on two public datasets, i.e., *IDIAP* and *CAVIAR*.

The *IDIAP* Head Pose dataset [59] (see Fig. 4(f)) comes from 8 meeting sequences of 360×288 frame resolution, where two individuals were captured while discussing about various topics in a 4-person dialogue scenario. The total number of different subjects captured is 15. They had their head orientations continuously annotated using a magnetic field location and orientation sensor tracker. The video repository has been lately employed for the CLEAR2007 head orientation estimation contest, following the protocol described in [34] (75×75 21152 samples were selected as training data and 23991 as testing data). Since the training samples are particularly biased on certain orientations, we flip them and then we randomly extract a subset of 5288 images, obtaining a balanced training pool. It represents a valuable benchmark set since the annotations express the pan, tilt and roll angles of the head pose. The best performances

3. i-LIDS dataset, <http://tna.europarchive.org/20100413151426/scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/i-lids/index.html>

are set in [60] (tilt) and [34] (pan, roll).

The CAVIAR dataset [61] (see Fig. 4(e)) is a more challenging set for the estimation task due to the low resolution of the images and the presence of occlusions. The considered head samples, resized to 50×50 pixels, come from a set of sequences which have 1500 frames on average, acquired from a real surveillance camera located in a shopping centre in Lisbon. The dataset is composed by two subsets: the first is made by non-occluded head images for a total number of 21325 examples (10660 as training and 10665 as testing set), the second consists in a dataset of 21691 images partitioned in 10802 training and 10889 for testing. For the best performance on this dataset, please see Sec. 5.4.

Finally, for the body orientation task, we introduce a novel dataset dubbed Human Orientation Classification (HOC) [56]. Even if this task has recently attracted the attention of researchers (see for example [42]) no public available datasets are present in the literature (except the ViPER dataset [62], but it has a very low number of elements, limited to 632). HOC (see Fig. 4(c)) is derived by the ETHZ [63] human re-acquisition dataset representing pedestrians in different orientations and (background) conditions, captured by hand-held cameras.

ETHZ is structured in three sequences for a total of 11881 images (6860 in the training and 5021 in the testing set), each image 64×32 pixels containing a pedestrian. We manually split the images into 4 orientation classes (Front, Back, Left, and Right), individuating a training and a testing partition. The dataset is complex because of the low resolution, severe illumination artifacts, occlusions and consistent scale changes.

5.2 Geometrical properties of Sym_d^+

The numerical evaluation of the curvature κ_{I_d} in correspondence of the samples of a particular dataset allows to understand how concave is the related region of Sym_d^+ . In Tab. 1, the mean value and the standard deviation of κ_{I_d} of 1K random elements for all the datasets are reported (note that QMUL[†] refers to the QMUL dataset with the background class). These values are calculated by considering each covariance matrix of WARCO as an independent sample, for all the WARCO descriptors of a single dataset.

In the same table, we calculate the mean values of the Frobenius distance $d_{\mathcal{E}}$ (27), the CBH1 distance d_{CBH1} (28) and the geodesic distance $d_{\mathcal{G}}$ (30) between all the possible couples of the above elements. In addition, we compute the mean error and its standard deviation, considering as Frobenius (CBH1) error the absolute value of the difference between a Frobenius (CBH1) distance value and the corresponding geodesic one.

Many observations can be drawn: first, larger patches seem to lie in flatter regions, and this assumption will be validated heuristically, in a more exhaustive fashion, later in the section. Considering the approximated



Fig. 4. Examples of the (a) HIIT, (b) HOCoffee, (c) HOC, (d) QMUL, (e) CAVIAR, and (f) IDIAP datasets used in the experimental part. In (a), (b), (c), and (d), each row correspond to a different class. In (e) and (f), head orientation is estimated by regression. Examples are ranked from the left to the right proportionally to their degree of difficulty.

distances, the approximation error in the case of large patches seems to be lower than in the case of tiny regions, and in particular, as expected, $Err_{d_{\mathcal{E}}} \leq Err_{d_{CBH1}}$; the same applies for the standard deviations of the errors. Finally, one can note that for the mean values the inequality $d_{\mathcal{E}} \leq d_{CBH1} \leq d_{\mathcal{G}}$ holds systematically.

5.3 Head Orientation Classification

QMUL Head dataset. We test our WARCO classifier adopting both the Frobenius and the CBH distances, against the template-based discriminative approach presented in [24] and the ARCO LogitBoost-based strategy [13], the latter being the current best approach. To reproduce the former method, we considered the image features provided by the dataset authors and we follow the same experimental protocol. The confusion matrices are reported in Fig. 5, considering 4 and 5 (4 orientations plus the background) classes. WARCO with CBH1 distances get the highest average classification scores. One should also pay attention to Fig. 5(h), where the accuracy in classifying the background class rises of about 10% with respect to the previous state-of-the-art results depicted in Fig. 5(f). This gap is due to the CBH distance: actually, background samples are located in zones with higher curvature (validated experimentally), far from I_d , so that the contribution given by the CBH

Dataset name	Dataset attrib.			WARCO		κ_{I_d}		d_ε		d_{CBH1}		d_G
	obj. of int.	# images	avg. obj. dim.	patches number	patch dim.	mean	standard dev.	mean	Err_{d_ε}	mean	$Err_{d_{CBH1}}$	mean
QMUL	head	16k	50×50	25	16×16	-0.035	0.017	7.78	0.99,(0.41)	8.21	0.60,(0.27)	8.78
QMUL †	head	20k	50×50	25	16×16	-0.038	0.020	8.65	0.98,(0.41)	9.13	0.59,(0.27)	9.65
HIIT	head	24k	50×50	25	16×16	-0.031	0.018	7.02	0.88,(0.42)	7.41	0.55,(0.28)	8.02
HOCoffee	head	18k	50×50	25	16×16	-0.035	0.015	6.40	0.88,(0.39)	8.37	0.52,(0.25)	8.88
CAVIAR (Cl.)	head	21k	50×50	25	16×16	-0.041	0.021	8.59	1.18,(0.41)	9.16	0.67,(0.26)	9.73
CAVIAR (Occ.)	head	22k	50×50	25	16×16	-0.043	0.026	8.12	1.19,(0.39)	8.88	0.69,(0.25)	9.12
IDIAP	head	66k	75×75	25	24×24	-0.014	0.006	4.79	0.43,(0.19)	5.01	0.27,(0.12)	5.34
HOC	human	11k	62×132	40	24×24	-0.024	0.014	7.67	0.59,(0.30)	7.99	0.37,(0.19)	8.41

TABLE 1

Curvature analysis and distance comparison of different datasets. κ_{I_d} , d_ε , d_{CBH1} , and d_G are compared on the same covariance representation (see Eq. (31)). The errors Err_{d_ε} and $Err_{d_{CBH1}}$ are shown with their mean value and its standard deviation (in parenthesis). See Sec. 5.2 for other details.

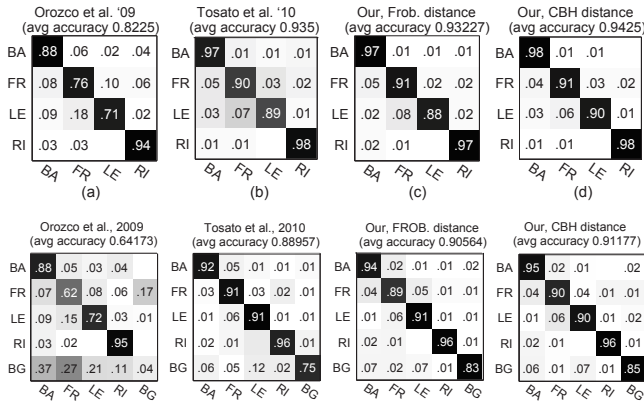


Fig. 5. Examples and statistics of the 4 and 5-class original dataset taken by Orozco et al. [24]. (a) and (e): the original results by Orozco et al. approach [24]. (b) and (f): the Tosato et al. approach [13]. (c), (d), (g), and (h): the proposed approach.

expansion becomes critical in better capturing the local geometry.

HOCoffee dataset.

In this case we have 6 orientations. In Fig. 6(e), the qualitative performances, and in Fig. 6(c) and (d), the quantitative performances are reported considering both Frobenious (FROB) d_ε distance (27) and the CBH1 d_{CBH1} distance.

HIIT dataset.

As one can note in Fig. 6(a) and (b), the performance of our framework are rather high, in fact, using d_{CBH1} to measure the distance among covariance matrices the average accuracy is 97%. This means that our classifier manages easily low resolution head images classifying the orientation precisely.

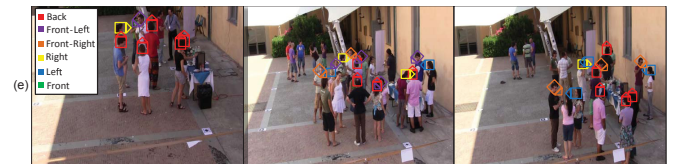
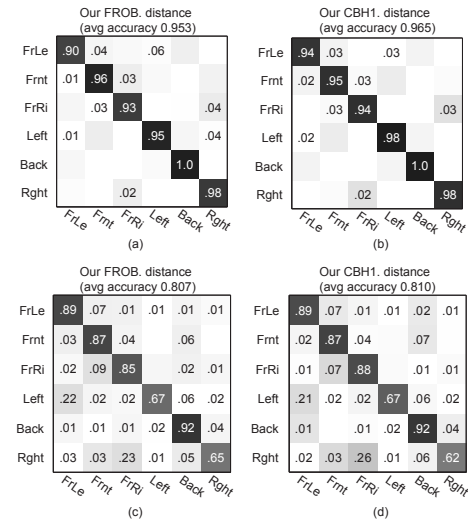


Fig. 6. Confusion matrices on the (a) and (b) HIIT, and (c) and (d) CoffeeBreak head orientation datasets [56]. (e) shows a qualitative result on the CoffeeBreak dataset.

5.4 Head Orientation Estimation by Regression.

In this context, we replace the SVM classifier with a SVM regressor [64].

IDIAP Head Pose. The head orientation evaluation protocol is taken from [34]: in each one of the 8 meetings of the test set, we have 1 minute of recording for the testing, for a total of 1500 test samples. We adopt the three error measures suggested by the protocol, which

	pan			tilt			roll		
	mean	std	med	mean	std	med	mean	std	med
Hist+Correlation [34]	16.2	13.6	13.1	22.4	15.0	19.1	15.1	12.0	12.5
Correlation+Shape [34]	19.0	17.4	14.2	26.4	17.5	21.5	16.1	12.7	13.4
Texture [34]	13.6	14.9	8.3	17.6	13.8	12.8	11.5	10.3	12.9
Texture+Color [34]	8.7	9.1	6.2	19.1	15.4	14.0	9.7	7.1	8.6
Neural Network [60]	14	—	—	9.2	—	—	—	—	—
Exemplar-based Tracking [65]	8.8	—	—	9.4	—	—	9.89	—	—
Large-margin paradigm [66]	12.5	15.8	—	8.5	11.3	—	8.5	9.6	—
Our, FROB. distance	10.90	10.75	7.87	4.81	5.98	2.93	4.65	4.22	3.80
Our, CBH1 distance	10.30	10.61	7.13	4.46	5.26	2.54	4.33	3.84	3.33

TABLE 2

Pan, tilt and roll error statistics over evaluation data of IDIAP dataset.

are the absolute differences with the ground-truth pan, tilt and roll angles. Table 2 summarizes our results considering all the methods in the literature that followed the protocol above. As one can observe, we reach good results concerning the pan, while we define the best scores with the tilt and roll angles. In Fig. 7, we report an analysis of the performances obtained by our framework *per sample*, on all the samples (employing the CBH1 distance), as compared to the ground-truth.

	pan		
	mean	std	med
Robertson & Reid [37]	76.4	55.8	70.1
WARCO, Clean, FROB. distance	22.65	18.44	17.09
WARCO, Clean, CBH1 distance	22.21	18.38	16.90
WARCO, Occluded, FROB. distance	36.90	25.23	31.73
WARCO, Occluded, CBH1 distance	35.26	24.58	30.70

TABLE 3

Pan error statistics over evaluation data of CAVIAR dataset for both non-occluded and occluded cases.

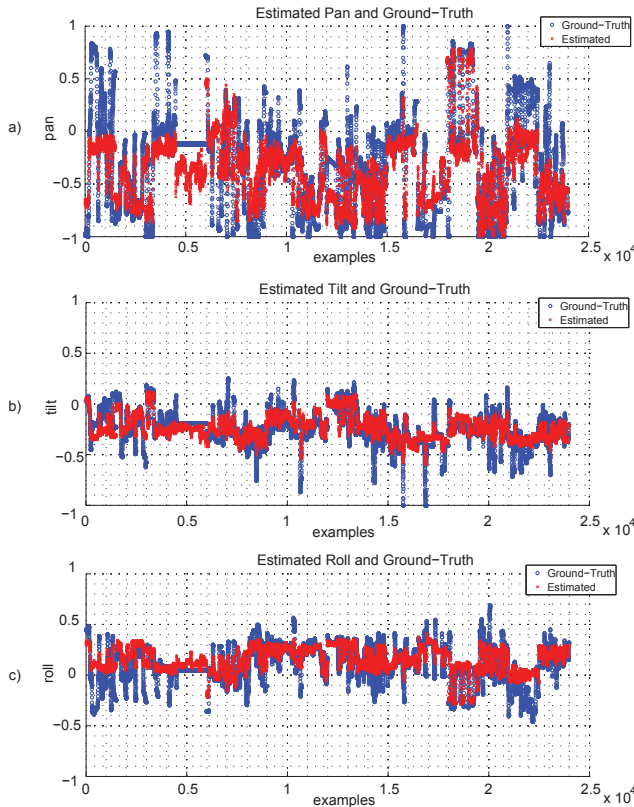


Fig. 7. Using our framework with the CBH1 distance, we show the difference between the estimated orientation and the ground-truth for pan (a), tilt (b), and roll (c) angles for the IDIAP head pose dataset (better viewed in colors).

CAVIAR. We consider the best competitor for this

dataset, which is the method presented in [37]. Unfortunately, we had difficulties in producing a fair comparison. In this paper [37], ground truth annotations are made by the authors, which unfortunately are not compatible with that provided together with the dataset. In practice, they represent a quantized version of the original annotations. Employing the original annotations, we individuate two datasets, one formed by non-occluded samples, the other with occlusions, and we estimate the pan angle on both sets. Results are shown in Table 3, where, as in [37], the mean, the standard deviation and the median of the errors are reported.

Two main considerations pop out. The first one is that our approach gets lower errors than [37]. Apart from the different methodologies in getting ground truth data, that should make the task of [37] easier than ours, WARCO is noticeably more accurate. The second observation is that the errors of WARCO in the occluded cases are not dramatically higher than the un-occluded cases, and this is due to the nature of WARCO, i.e., an ensemble of local classifiers.

5.5 Human Orientation Classification dataset.

In this case, WARCO is computed on 40 overlapped patches of 24×24 pixels. In Fig. 8 one can see the accuracy result achieved by our algorithm. Despite the heavy occlusions and the bad illumination conditions, the average accuracy reaches 79%. It is worth noting how the Front and the Back classes are nicely separated: this is an impressive results, since here the most noticeable difference between the two classes lies in the head portion, which is relatively small.

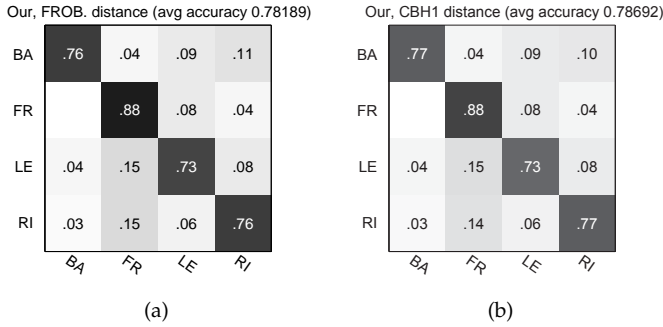


Fig. 8. Confusion matrices showing the performances between the WARCO method using (a) the Frobenius distance ($d_{\mathcal{E}}$) and (b) the CBH1 distance (d_{CBH1}).

5.6 Scale issues.

Here, we stress the capability of WARCO of working at low resolution, and we explore the relation between patch dimensions and manifold curvature κ_{I_d} . We produce two additional experimental sessions, where we reduce the image dimensions of each dataset by a factor of 0.75 and 0.5. Consequently, we reduce by the same factor the architecture of WARCO. In Fig. 9, we report the results concerning the classification, and in Fig. 10 we show the results for the regression task.

As one can note, the smaller the size of the object, the higher the curvature. Furthermore, it is valuable to observe how the CBH1 distance-based framework behaves with respect to the Frobenius distance-based technique at the different resolutions: the lower the resolution, the bigger the gap between CBH1 and the Frobenius-based strategy. Once again, this demonstrates that the contribution of CBH1 is in general more helpful in highly curved manifold regions.

6 CONCLUSIONS

We presented a method for characterizing tiny images of pedestrians in a surveillance scenario, specifically, for performing head orientation and body orientation estimation, employing arrays of covariances as descriptors, named WARCO. Given the results achieved, we are confident that the framework will be adopted as standard tool in surveillance applications.

We also think that our work is valuable beyond the scope of the contingent application: actually, we suggested a theoretically sound way to deal with covariance matrices, like they were points lying in an Euclidean space. This came with a measure for approximating geodesic distances, the CBH1 measure, that works better than standard Euclidean distance.

Future research on this topic will check whether the triangular inequality holds for CBH1, in order to validate CBH1 as genuine distance. Furthermore, we plan to extend WARCO as an action descriptor, including the temporal dimension in the analysis, and to inject multiple kernel reasoning for learning the weights of WARCO, instead of calculating them independently.

REFERENCES

- [1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1743–1759, 2009.
- [2] K. Smith, S. Ba, J. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE PAMI*, vol. 30, no. 7, pp. 1–18, 2008.
- [3] N. Robertson and I. Reid, "Automatic reasoning about causal events in surveillance video," *EURASIP Journal on Image and Video Processing*, 2011.
- [4] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, A. D. Bue, D. Tosato, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of f-formations," in *Proceedings of British Machine Vision Conference*, 2011.
- [5] M. Cristani, A. Pesarin, A. Vinciarelli, M. Crocco, and V. Murino, "Look at who's talking: Voice activity detection by automated gesture analysis," in *Proceedings of the Workshop on Interactive Human Behavior Analysis in Open or Public Spaces (InterHub 2011)*, 2011.
- [6] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Trans. PAMI*, vol. 30, no. 10, pp. 1713–1727, 2008.
- [7] —, "Region covariance: A fast descriptor for detection and classification," in *Proc. ECCV*, 2006, pp. II: 589–600.
- [8] M. Donoser and H. Bischof, "Using covariance matrices for unsupervised texture segmentation," in *ICPR*, 2008, pp. 1–4.
- [9] J. Yao and J. Odobez, "Fast Human Detection from Videos Using Covariance Features," in *The Eighth International Workshop on Visual Surveillance*, 2008.
- [10] B. Wu and R. Nevatia, "Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection," in *Proc. CVPR*, 2008.
- [11] H. Karcher, "Riemannian Center of Mass and Mollifier Smoothing," *Comm. Pure and Applied Math.*, vol. 30, pp. 509–541, 1997.
- [12] J. Duistermaat and J. Kolk, *Lie groups*. Springer Verlag, 2000.
- [13] D. Tosato, M. Farenzena, M. Cristani, M. Spera, and V. Murino, "Multi-class classification on riemannian manifolds for video surveillance," in *Proc. ECCV*. Springer, 2010, pp. 378–391.
- [14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, vol. 1, 2005, p. 886.
- [15] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," in *Proc. CVPR*. IEEE, 2010, pp. 3539–3546.
- [16] J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *Proc. CVPR*. IEEE, 2009.
- [17] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. ECCV*. Springer, 2006, pp. 589–600.
- [18] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.
- [19] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *IJCV*, vol. 82, no. 2, pp. 185–204, 2009.
- [20] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. BMVC*, 2009.
- [21] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. PAMI*, vol. 31, pp. 2179–2195, 2009.
- [22] J. Odobez and S. Ba, "A cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose," in *Proc. ICME*. IEEE, 2007, pp. 1379–1382.
- [23] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. CVPR*, vol. 1, 2009, pp. 1014–1021.
- [24] J. Orozco, S. Gong, and T. Xiang, "Head pose classification in Crowded Scenes," in *Proc. BMVC*, 2009.
- [25] P. Sabzmejdani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proc. CVPR*, 2007, pp. 1–8.
- [26] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *Proc. CVPR*. IEEE, 2008, pp. 1–8.
- [27] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," *Pattern Recognition*, vol. I, pp. 82–91, 2008.

Dataset	Obj. Size	κ_{I_d}	$d_{\mathcal{E}}$ avg acc.	d_{CBH1} avg acc.	Dataset	Obj. Size	κ_{I_d}	$d_{\mathcal{E}}$ avg acc.	d_{CBH1} avg acc.
QMUL	25×25	-0.0509	78%	80%	QMUL †	25×25	-0.0571	74%	76%
	38×38	-0.0448	89%	90%		38×38	-0.0470	86%	87%
	50×50	-0.0361	91%	92%		50×50	-0.0345	90%	91%

(a) (b)

Dataset	Obj. Size	κ_{I_d}	$d_{\mathcal{E}}$ avg acc.	d_{CBH1} avg acc.	Dataset	Obj. Size	κ_{I_d}	$d_{\mathcal{E}}$ avg acc.	d_{CBH1} avg acc.
HIIT	25×25	-0.0571	88%	90%	HOCoffe	25×25	-0.607	62%	66%
	38×38	-0.0571	95%	96%		38×38	-0.0430	78%	80%
	50×50	-0.0571	96%	96%		50×50	-0.0345	80%	80%

(c) (d)

Dataset	Obj. Size	κ_{I_d}	$d_{\mathcal{E}}$ avg acc.	d_{CBH1} avg acc.
HOC	66×31	-0.0320	71%	73%
	99×47	-0.0230	77%	78%
	132×62	-0.0192	78%	78%

(e)

Fig. 9. Comparative study of the performances of the proposed statistical classification framework.

Dataset	Obj. Size	κ_{I_d}	Avg Pan Err.		Dataset	Obj. Size	κ_{I_d}	Avg Pan Err.	
CAVIAR (Clean)	25×25	-0.0437	$d_{\mathcal{E}}$	d_{CBH1}	CAVIAR (Occluded)	25×25	-0.045	$d_{\mathcal{E}}$	d_{CBH1}
	38×38	-0.0426	27.15	25.63		38×38	-0.044	41.00	38.00
	50×50	-0.0415	22.65	21.58		50×50	-0.043	37.00	36.33
			19.74	19.73				36.90	35.26

(a) (b)

Dataset	Obj. Size	κ_{I_d}	Avg Pan Err.		Avg Tilt Err.		Avg Roll Err.	
IDIAP	38×38	-0.0293	$d_{\mathcal{E}}$	d_{CBH1}	$d_{\mathcal{E}}$	d_{CBH1}	$d_{\mathcal{E}}$	d_{CBH1}
	56×56	-0.0175	16.18	16.07	6.67	6.47	5.02	4.97
	75×75	-0.0143	12.35	12.03	5.18	5.01	4.94	4.82
			10.90	10.30	4.81	4.46	4.65	4.33

(c)

Fig. 10. Comparative study of the performances of the proposed statistical regression framework.

- [28] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. PAMI*, vol. 31, no. 4, pp. 607–626, 2009.
- [29] X. Wang, T. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. ICCV*. IEEE, 2010, pp. 32–39.
- [30] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. CVPR*. IEEE, 2010, pp. 1030–1037.
- [31] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, "Part-based feature synthesis for human detection," *Proc. ECCV*, vol. I, pp. 127–142, 2010.
- [32] Z. Lin and L. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Proc. ECCV*. Springer, 2008, pp. 423–436.
- [33] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. PAMI*, vol. 31, no. 4, pp. 607–626, 2009.
- [34] S. Ba and J. Odobez, "Evaluation of multiple cue head pose estimation algorithms in natural environments," in *Proc. ICME*. IEEE, 2005, pp. 1330–1333.
- [35] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Proc. CVPR*, 2011.
- [36] D. Huang, M. Storer, F. De la Torre, and H. Bischof, "Supervised local subspace learning for continuous head pose estimation," in *Proc. CVPR*, 2011.
- [37] N. Robertson and I. Reid, "Estimating gaze direction from low-resolution faces in video," in *Proc. ECCV*. Springer, 2006, pp. 402–415.
- [38] A. Agarwal and B. Triggs, "A Local Basis Representation for Estimating Human Pose from Cluttered Images," in *Proc. ACCV*. Springer, 2006, pp. 50–59.
- [39] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Proc. CVPR*. IEEE, 2009, pp. 1365–1372.
- [40] D. Tran and D. Forsyth, "Improved Human Parsing with a Full Relational Model," in *Proc. ECCV*. Springer, 2010, pp. 227–240.
- [41] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human Detection Using Partial Least Squares Analysis," in *Proc. ICCV*, 2009.
- [42] M. Enzweiler and D. Gavrilu, "Integrated pedestrian classification and orientation estimation," in *Proc. CVPR*. IEEE, 2010, pp. 982–989.
- [43] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, vol. 2, 2006, pp. 2169–2178.
- [44] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Efficient similarity search for covariance matrices via the jensen-bregman logdet divergence," *Computer Vision, IEEE International Conference on*, vol. 0, pp. 2399–2406, 2011.
- [45] P. Fletcher, C. Lu, S. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," in *Proc. ECCV Workshop*, vol. 23, no. 8. IEEE, 2004, pp. 995–1005.
- [46] P. Fillard, X. Pennec, V. Arsigny, and N. Ayache, "Clinical DT-MRI estimation, smoothing, and fiber tracking with log-Euclidean metrics," *IEEE Trans. MI*, vol. 26, no. 11, pp. 1472–1482, 2007.
- [47] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 1, p. 328, 2008.
- [48] P. Fletcher, S. Venkatasubramanian, and S. Joshi, "Robust statistics on Riemannian manifolds via the geometric median," in *Proc. CVPR*. IEEE, 2008, pp. 1–8.

- [49] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen, "Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations," in *Proc. ECCV*. Springer, 2010, pp. 43–56.
- [50] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Efficient similarity search for covariance matrices via the jensen-bregman logdet divergence," in *Proc. IICV*. IEEE, 2011, pp. 2399–2406.
- [51] I. Chavel, *Riemannian Geometry - A modern introduction*. Cambridge University Press, Cambridge, 2006.
- [52] E. Sernesi, *Linear algebra: a geometric approach*. Chapman & Hall/CRC, 1993.
- [53] L. Saul and M. Jordan, "Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones," *Machine Learning*, vol. 37, no. 1, pp. 75–87, 1999.
- [54] J. Zhang, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, p. 2007, 2007.
- [55] P. Dollár, "Piotr's Image and Video Matlab Toolbox (PMT)," <http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html>.
- [56] D. Tosato, "ARCO (array of covariance matrices), code and datasets," <http://sites.google.com/site/diegotosato/ARCO>.
- [57] N. Gourier, "Head pose image database (pointing'04 icpr workshop)," <http://www-prima.imag.fr/Pointing04/data-face.html>.
- [58] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "The cmu multi-pose, illumination, and expression (multi-pie) face database," Technical report, Carnegie Mellon University Robotics Institute. TR-07-08, Tech. Rep., 2007.
- [59] J. M. Odobez, "IDIAP head pose database," <http://www.idiap.ch/dataset/headpose>.
- [60] M. Voit, K. Nickel, and R. Stiefelhagen, "Head Pose Estimation in Single- and Multi-view Environments - Results on the CLEAR'07 Benchmarks," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2008, vol. 4625, ch. 29, pp. 307–316.
- [61] R. Fisher, "CAVIAR case scenarios," <http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/>.
- [62] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition and tracking," in *PETS*, 2007.
- [63] W. R. Schwartz, "ETHZ dataset for appearance-based modeling," <http://www.liv.ic.unicamp.br/~wschwartz/datasets.html>.
- [64] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [65] S. Ba and J.-M. Odobez, "From camera head pose to 3d global room head pose using multiple camera views," in *Int'l. Workshop Classification of Events Activities and Relationships (CLEAR 07)*, ser. Lecture Notes in Computer Science. Springer, 2007.
- [66] E. Ricci and J.-M. Odobez, "Learning large margin likelihoods for realtime head pose tracking," in *Proceedings of the 16th IEEE international conference on Image processing*, ser. ICIP'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 2565–2568. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1819298.1819450>



Diego Tosato received the Laurea degree in Information Technology and Laurea Specialistica degree in Intelligent and Multimedia Systems from the University of Verona, Italy, in 2006 and 2008. He is currently a Ph.D. student in the Department of Computer Science, University of Verona working with the Vision, Image Processing and Sounds (VIPS) Lab. His research interests are in computer vision, machine learning, and statistical methods to image understanding problems.



national and international conferences.

Mauro Spera is currently Associate Professor of Geometry at the Computer Science Department of Faculty of Sciences of the University of Verona, Italy. His present research interests include infinite dimensional differential geometry, geometric methods in quantum mechanics, vortex theory and link invariants, loop space extensions of the index theory, geometrical aspects of computer vision. He authored/coauthored more than forty research papers on internationally renowned journals and participated in several



fusion techniques, with applications on surveillance, segmentation, and image and video retrieval.

Marco Cristani received the Laurea degree in 2002 and the Ph.D. degree in 2006, both in computer science from the University of Verona, Verona, Italy. He is now Assistant Professor with the Dipartimento di Informatica, University of Verona, working with the VIPS Lab. He is also Team Leader with the Istituto Italiano di Tecnologia, Genova, working with the PAVIS Lab. His main research interests include statistical pattern recognition, generative modeling via graphical models, and nonparametric data



learning, in particular, probabilistic techniques for image and video processing, with applications on video surveillance, biomedical image analysis and bioinformatics. He is also member of the editorial board of Pattern Recognition, Pattern Analysis and Applications, and Machine Vision & Applications journals, as well as of the IEEE Transactions on Systems Man, and Cybernetics - Part B: Cybernetics. Finally, prof. Murino is senior member of the IEEE and Fellow of the IAPR.

Vittorio Murino is full professor and head of the PAVIS department (Pattern Analysis and Computer Vision) at the Istituto Italiano di Tecnologia, Genova, Italy. He received the Ph.D. in Electronic Engineering and Computer Science in 1993 at the University of Genova, Italy. Then, he was first at the University of Udine and, since 1998, at the University of Verona, where he served as chairman of the Department of Computer Science from 2001 to 2007. His research interests are in computer vision and machine